Weir, C. J., Heazell, A., Whyte, S., & Norman, J. E. (2020). Evaluating improvement interventions using routine data to support a learning health system: research design, data access, analysis and reporting. *BMJ Quality and Safety*. https://doi.org/10.1136/bmjqs-2019-010068

Evaluating improvement interventions using routine data to support a learning health system: research design, data access, analysis and reporting

Running title: Improvement intervention evaluations

*Christopher J Weir PhD[1], Alexander E P Heazell PhD[2,3], Sonia Whyte MSc[4], Jane E Norman MD[4,5]

[1] Edinburgh Clinical Trials Unit, Centre for Population Health Sciences, Usher Institute of Population Health Sciences and Informatics, the University of Edinburgh, Nine Edinburgh BioQuarter, 9 Little France Road, Edinburgh, EH16 4UX, UK

[2] Tommy's Maternal and Fetal Health Research Centre, School of Medical Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

[3] St. Mary's Hospital, Central Manchester University Hospitals NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK

[4] Tommy's Centre for Maternal and Fetal Health, MRC Centre for Reproductive Health, Queen's Medical Research Institute, Edinburgh, EH16 4TY, UK

[5] Faculty of Health Sciences, University of Bristol, Bristol, UK.

* Corresponding author, Christopher.Weir@ed.ac.uk, telephone 0131 651 9957

Contribution to authorship

CJW drafted this Viewpoint; CJW, AEPH, SW and JEN critically reviewed and edited its content.

Disclosure of interests

All authors were collaborators on the AFFIRM clinical trial which is cited in this Commentary; AEPH was a lead investigator on the SPIRE study.

Funding

CJW was funded in this work by NHS Lothian via the Edinburgh Clinical Trials Unit. The AFFIRM study was investigator-initiated and funded by the Chief Scientist Office, Scottish Government (CZH/4/882), Tommy's, and Sands. SPIRE was funded by NHS England.

Word count 2376

Details of ethics approval

AFFIRM received a favourable opinion from Scotland A Research Ethics Committee (Ref 13/SS/0001)

SPIRE received a favourable opinion from West Midlands, Edgbaston Research Ethics Committee (Ref 17/WM/0197).

Introduction: learning health systems

Friedman and colleagues[1] outline a vision of the learning health system, founded on the sharing of data and achieved through alignment of information technology, advanced analytics and clinical expertise. The Institute of Medicine[2] recognises the potential of the learning health system to generate new information automatically during the delivery of healthcare, offering continual opportunities to improve healthcare processes to the benefit of public health. The learning health system is promoted as a mechanism to accelerate the adoption of effective treatments into clinical practice, shortening the extended delay[3] from publication of research findings to implementation. Furthermore, it harbours the ambition to deliver personalised medicine to each service user, rather than the systematic provision of identical care to groups of patients who share the same characteristics. Worldwide escalating costs in healthcare provision due to demographic changes, compounded by ongoing use of ineffective tests and treatments, make it critically important to harness the efficiency gains of a learning health system. Of its many potential characteristics, one distinct attribute of the learning health system is to enable efficient investigation of whether strategies promoting implementation of best practice (such as educational initiatives or care bundles) actually work. A systematic review identifying a low frequency of reports on the evaluation and impact of learning health systems[4] prompted us here to reconsider the central requirements for evaluation of improvement interventions within the learning health system.

Here, we reflect on two recent implementation studies, both utilising efficiencies of the learning health system (minimising data acquisition and relying heavily on data acquisition from existing medical records), to illustrate the key issues arising when incorporating this type of research into the learning health system. In this paper we argue that the presence of infrastructure which facilitates data sharing, combined with appropriate research design, analysis and reporting, are essential elements in the evaluation of healthcare improvement interventions.

Data, research design and the learning health system

Ready access to routine data in the learning health system is a priority for commissioners of implementation research investigating system-wide healthcare interventions and researchers performing those evaluations. In turn, this becomes an important consideration for those configuring new or updated electronic health record systems. Here, using both the AFFIRM trial (Can Promoting Awareness of Fetal movements and Focussing Interventions Reduce fetal Mortality - a stepped wedge cluster randomised trial)[5] and the SPIRE project (Saving Babies' Lives Project Impact and Results Evaluation)[6] as case studies, we elucidate key challenges in three areas central to the conduct of such research at scale: data access; research design; and analysis and reporting. We conclude with recommendations for research commissioners, researchers and managers of healthcare information systems to enable the benefits of a learning health system in evaluating improvement interventions to be achieved.

AFFIRM and SPIRE

The AFFIRM trial [5] evaluated a package of care aimed at reducing stillbirth. The complex intervention evaluated in AFFIRM (in over 400,000 women) combined strategies for increasing awareness among pregnant women of the need to report decreased fetal movements (DFM) promptly, with a structured management plan to identify fetal compromise and achieve timely delivery in suspected and confirmed cases of DFM. SPIRE [6] aimed to determine whether the Saving Babies Lives Care Bundle reduced the occurrence of stillbirth by applying best practice to four components of maternity care. Its evaluation included more than 95,000 deliveries per year in 19 secondary and tertiary maternity units covering 9 local authorities in England. Both AFFIRM and SPIRE were pragmatic trials, applying "real world" implementation at scale, and evaluating outcomes through the use of routinely collected data.

Data access

The first challenge for studies such as AFFIRM and SPIRE arises from the use of routinely collected data. The research team relies on the required measurements being available and accurately

recorded. For AFFIRM, the Scottish Birth Record, Maternal Inpatient and Day Case Records contained all the necessary data for sites in Scotland on inpatient care, mother and baby characteristics, and pregnancy and baby outcomes. Substantial data management input was then needed to map separate data sources from study sites in England, Wales, Northern Ireland and the Republic of Ireland onto a common database (409,175 women in total) stored securely in the safe haven at National Health Service (NHS) National Services Scotland. Not all desirable information was routinely available; for example the frequency with which investigations (such as cardiotocography) were used (to assess fidelity of implementation of the intervention) and the results of those examinations had to be gathered separately in site audits.

The quality of routinely collected data is further complicated by variation in nature of recording and coding of individual activities e.g. ultrasound scan to measure fetal growth, such that in SPIRE only the overall average number of ultrasound scans performed could be calculated rather than just those relevant to the area of study. These limitations to the nature and scope of routinely collected data disproportionately affect process measures (for instance, the number of women attending with DFM) as these are less likely to be recorded than outcomes (for example, stillbirth). As a result, intervention fidelity can be measured at an aggregate level (ward or healthcare facility) but not for individual patients.[7] In some instances, dependent on the nature of the intervention being evaluated, such cluster-level assessment of fidelity may be all that is required. Importantly, clinical coding dictionaries often do not include items for the presence of significant symptoms such as DFM. The fundamental importance of data management issues is reflected in the RECORD[8] reporting guidance for research using routinely collected data.

Access permissions are also key: in the European Union, the General Data Protection Regulation[9] incorporates a “legitimate interest” criterion which is currently being applied to support the use, in the presence of appropriate information governance safeguards, of routine data for research. Ease of

access to relevant and interpretable routine data is a fundamental requirement of any learning health system.

Research design

The second feature concerns the importance of advance planning of study design in the evaluation of improvement interventions. Prospective planning of the implementation and evaluation of the intervention enabled AFFIRM to set up a stepped-wedge cluster-randomised trial design[10] whereas SPIRE took the form of a natural experiment as its evaluation was only planned after the development and introduction of the intervention in early adopter sites.

Stepped-wedge design trials (Figure 1) commence with all of the clusters delivering treatment as usual. At regular pre-specified intervals each cluster (or group of clusters) implements the intervention and maintains this for the remainder of the trial. By the final time interval all clusters have adopted the intervention. Randomisation determines the time point at which each cluster commences the intervention.

Stepped-wedge designs are conceptually useful in enabling an empirical randomised evaluation of an intervention which is intended to be rolled out across an entire health system. Non-randomised designs, based on real-world evidence alone, provide an alternative evaluation framework. While some advocate the use of causal inference in non-randomised designs, consensus is lacking and approaches such as propensity scoring to address confounding require sensitivity analyses to provide further assurance on their validity.[11] The confounding of intervention effects with time in non-randomised simple before-and-after designs[10] is also present in stepped-wedge designs. Time effects may be adjusted for in the analysis of stepped-wedge trials to reduce such confounding, although this requires the reasonably strong assumption that the underlying time trend is the same for all clusters. Interrupted time series analysis may be applied to observational study designs; while this offers another candidate approach to evaluating service-level interventions, it is also vulnerable to confounding. Furthermore it assumes a fixed time point at which change happens: as was found in

SPIRE,[6] this may not reflect the reality of introducing a complex intervention at service level. The key advantage of the stepped-wedge randomised design over non-randomised before-and-after designs is that it enables a contemporary comparison between clusters which have, and have not, implemented the intervention.

All designs which randomise clusters rather than individual participants incur a penalty which inflates the sample size required. It is notable that for a given number of participants per cluster, stepped-wedge designs require a smaller number of clusters (are more efficient) than parallel group cluster trials; however, this greater efficiency is lost where the within-cluster correlation of outcomes (the "intra-cluster correlation coefficient") is low.[12]

Depending on the nature and risk profile of the intervention being investigated, such cluster trials may not require individual informed consent from participants,[13] which potentially enhances study efficiency and representativeness of the trial population. Nevertheless, recommendations on informed consent in the Ottawa statement[14] emphasise that consent should be sought unless a waiver is clearly justified. Technical understanding of sample size requirements[15] and optimal design configurations[16] has developed for stepped-wedge trials.

Practical challenges include factors outside the control of the researcher, for example the closure of study sites; and, as encountered in AFFIRM, the dropout of sites, post-randomisation, for reasons such as the perceived costs of the intervention being studied. Sites must also be willing to agree to the intervention at baseline and to defer implementation until the time point allocated in the randomisation sequence – mistimed implementation adversely impacts on study integrity.[17] Such challenges emphasise the need to consider evaluation strategies concurrently with the development of interventions so that potential challenges can be identified and addressed.

It is also important for understanding of the service-level impact of an intervention to study the varying degree to which sites implement the intervention. Quantitatively, we can establish the level of fidelity to the intervention in individual study sites. Qualitative studies enable further insight using

contextual information on the barriers to and facilitators of adoption of a complex intervention at scale.

Evaluation designs which appropriately address the above considerations of precision (sample size), bias (time confounding) and the influence of mediating and moderating factors on intervention effectiveness are essential to facilitate learning in a health system.

Analysis and reporting

Finally, issues arise relating to statistical analysis and reporting. In AFFIRM, the stillbirth primary outcome was a rare event, with an expected frequency of about 0.44%. This, together with the multi-level models used to analyse data from a stepped-wedge design, required a sample size of several hundred thousand. Indeed, for some research questions several million participants will be needed.[18] Another statistical challenge when using routine data, which cannot be subjected to conventional clinical trial data querying, is missing data. The missing data handling techniques being used should be pre-specified and the further assumptions they make will need to be justified and tested in sensitivity analyses.

Two further statistical issues arise due to features of the stepped-wedge design. First, the confounding between time and intervention effect requires adjustment for secular trends, using assumptions such as similarity of time effects across clusters which may be difficult to verify empirically. Secondly, the intra-cluster correlation coefficient may vary over time or differ between treatment as usual and intervention. While all of these statistical issues can be accommodated in the analysis, each in turn adds complexity to the model and requires a further layer of assumptions to be made and verified.

Final reporting also requires careful consideration for trial designs where informed consent has not been sought. For analyses of routine data performed within a secure safe haven environment, disclosure checking must be applied[19], to ensure none of the resulting statistical outputs (tables, graphics or regression modelling) would potentially identify an individual.

The final requirements for evaluation of improvement interventions in the learning health system are therefore the tailoring of statistical analysis to be fit for purpose; and a reporting process which protects privacy in order to maintain trust in the use of routine healthcare data to inform service improvements.

Recommendations

Notwithstanding the challenges outlined above, the benefits of using routine data for evaluation of improvement as part of a learning healthcare system vastly outweigh the drawbacks. Healthcare technologies and needs are continually evolving, and the cost of healthcare ever spiralling upwards: to fail to innovate or to innovate without proper evaluation is at best lazy and at worst unethical. A learning healthcare approach allows evaluation at scale, for modest costs. Based on our experience with AFFIRM and SPIRE, we suggest some recommendations for clinical data champions, for researchers and commissioners of research, and for policy makers. Our recommendations cover routine data access, study design, and statistical analysis and reporting:

Accessing routinely collected data Those configuring or updating healthcare information systems should strive to establish unified electronic health records where such a facility is currently absent, since clinical audit is time consuming and resource intensive compared to directly accessing data via electronic health records. For example, in SPIRE 1,658 case notes were audited, 2,230 women responded to a questionnaire and 1,064 health professionals completed a survey. It is also vital that information governance of electronic health records systems should facilitate secure access to data for research and incorporate a process for timely disclosure checking of research outputs to protect the anonymity of patients.

Research study design Commissioners of improvement research should endorse the use of randomised designs such as the stepped-wedge trial to enable empirical evaluation of healthcare system-wide interventions. Given the likely variations in implementation of an intervention across

different sites, the randomised trial should be supported by a quantitative and qualitative process evaluation.

Statistical analysis and reporting Researchers conducting trials using routinely collected data should incorporate a pre-specified and fully justified plan of the missing data handling methods to be used, which will require appraisal of the methods of data collection and development of an understanding of possible reasons why each data item might be missing. Robustness of the statistical model being fitted should be verified by testing its assumptions in sensitivity analyses. Researchers should also structure their data management and reporting to take account of the RECORD guidance[8] in order to support research transparency and optimise the interpretability and reproducibility of the findings.

Conclusions

As shown in AFFIRM and other studies using remote follow-up via electronic health records, the conduct of efficient randomised evaluations of interventions at scale generates robust evidence to support improvement in a learning health system. Such trials depend on appropriate infrastructure (safe haven access to routine data), study design (stepped-wedge trial) and analysis and reporting methods. Together, these have the potential to enable society to realise the benefits of a learning health system.

References

1 Friedman C, Rubin J, Brown J, *et al.* Toward a science of learning systems: a research agenda for the high-functioning Learning Health System. *J Am Med Inf Assoc* 2015; **22**: 43–50.

2 Institute of Medicine. Digital infrastructure for the learning health system: The Foundation for Continuous Improvement in Health and Health Care. Workshop Series Summary. Washington DC: The National Academies Press, 2011.

3 Slote Morris Z, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. *J R Soc Med* 2011: **104**: 510–520.

4 Budrionis A, Gustav Bellika J. The Learning Healthcare System: Where are we now? A systematic review. *J Biomed Inform* 2016; **64**: 87-92.

5 Norman JE, Heazell AEP, Rodriguez A, *et al.* Awareness of fetal movements and care package to reduce fetal mortality (AFFIRM): a stepped wedge, cluster-randomised trial. *Lancet* 2018; **392**: 1629–38.

6 Widdows K, Reid HE, Roberts SA, Camacho EM, Heazell AEP. Saving babies' lives project impact and results evaluation (SPiRE): a mixed methodology study. *BMC Pregnancy Childbirth* 2018; **18**: 43.

7 Widdows K, Roberts SA, Camacho EM, Heazell AEP. Evaluation of the implementation of the Saving Babies' Lives Care Bundle in early adopter NHS Trusts in England. Maternal and Fetal Health Research Centre, Manchester, UK: University of Manchester, 2018. https://www.manchester.ac.uk/discover/news/download/573936/evaluationoftheimplementationofthesavingbabieslivescarebundleinearlyadopternhstrustsinenglandjuly2018-2.pdf (accessed 31 October, 2019).

8 Benchimol EI, Smeeth L, Guttmann A, *et al.* The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) Statement. *PLoS Med* 2015; **12**: e1001885.

9 https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed July 5, 2019).

10 Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *BMJ* 2015; **350**: h391.

11 Streiner DL, Norman GR. The pros and cons of propensity scores. *Chest* 2012; **142**: 1380-1382.

12 Baio G, Copas A, Ambler G, Hargreaves J, Beard E, Omar RZ. Sample size calculation for a stepped-wedge trial. *Trials* 2015; **16**: 354.

13 Ford I, Norrie J. Pragmatic trials. *N Engl J Med* 2016; **375**: 454–63.

14 Weijer C, Grimshaw JM, Eccles MP, *et al*. The Ottawa statement on the ethical design and conduct of cluster Randomized Trials. *PLoS Med* 2012; **9**: e1001346.

15 Hemming K, Girling A. A menu-driven facility for power and detectable-difference calculations in stepped-wedge cluster-randomized trials. *Stata J* 2014; **14**: 363–80.

16 Hemming K, Lilford R, Girling AJ. Hemming K, Lilford R, Girling AJ. Stepped-wedge cluster randomised controlled trials: a generic framework including parallel and multiple level designs. Statist. Med. 2015, 34 181–196. *Stat Med* 2015; **34**: 181–96.

17 Heim N, van Stel HF, Ettema RG, van der Mast RC, Inouye SK, Schuurmans MJ. HELP! Problems in executing a pragmatic, randomized, stepped wedge trial on the Hospital Elder Life Program to prevent delirium in older patients. *Trials* 2017; **18**: 220.

18 Tannen R, Weiner M, Xie D. Use of primary care electronic medical record database in drug efficacy research on cardiovascular outcomes: comparison of database and randomised controlled trial findings. *BMJ* 2009; **338**: b81.

19 Hundepool A, Domingo-Ferrer J, Franconi L, *et al.* Statistical Disclosure Control, First Edition. Chichester: Wiley, 2012.

Figure 1 The stepped-wedge trial design

Example stepped-wedge trial design. Initially (month 1) no clusters have implemented the intervention. In month 2 a cluster (number 4) is randomly selected to implement the intervention. This process continues, one randomly selected cluster implementing the intervention each month, until by the end of the trial all eight clusters have adopted the intervention.